# A Survey on Model Compression for Large Language Models

*Year: 2023* **|** *Citations: 331* **|** *Authors: Xunyu Zhu, Jian Li, Yong Liu*

## Abstract

Abstract Large Language Models (LLMs) have transformed natural language processing tasks successfully. Yet, their large size and high computational needs pose challenges for practical use, especially in resource-limited settings. Model compression has emerged as a key research area to address these challenges. This paper presents a survey of model compression techniques for LLMs. We cover methods like quantization, pruning, and knowledge distillation, highlighting recent advancements. We also discuss benchmarking strategies and evaluation metrics crucial for assessing compressed LLMs. This survey offers valuable insights for researchers and practitioners, aiming to enhance efficiency and real-world applicability of LLMs while laying a foundation for future advancements.